Pr3_Anàlisis_dades_titanic

Ricard Deza Tripiana

7 de gener, 2019

Contents

1. Descripció del dataset

Perquè és important i quina pregunta/problema pretén respondre?

El dataset que tractarem en aquesta pràctica contè les dades dels passatger del Titanic. Com ja sabem, el Titanic va ser una embarcació transatlàntica la qual va patir un accident i naufragà en el seu viatge de inauguració.

L'objectiu d'aquesta pràctica és la neteja, tractament i anàlisis de les dades per tal de poder respondre a la pregunta sobre si existeix algun grup de passatgers amb més probabilitats de sobreviure a l'accident.

En primer lloc, per tal de poder analitzar el conjunt de dades, llegirem els fitxers proporcionats. Disposem de dos fitxers, train i test. A partir d'ara sempre parlarem del conjunt de dades d'entrenament. En el cas que es tracti del conjunt de prova, ho especificarem.

```
# Lectura del fitxer en un dataframe (train)
passatgers <- read.csv(paste(ruta, "train.csv", sep = ""), header = TRUE, sep = ",",
    na.strings = "NA", encoding = "UTF-8")
# Lectura del fitxer en un dataframe (test)
passatgers_test <- read.csv(paste(ruta, "test.csv", sep = ""), header = TRUE, sep = ",",
    na.strings = "NA", encoding = "UTF-8")
# Primers registres del datatset
kable(head(passatgers[, 1:6]))
```

PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
3	1	3	Heikkinen, Miss. Laina	female	26
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
5	0	3	Allen, Mr. William Henry	male	35
6	0	3	Moran, Mr. James	male	NA

```
head(passatgers[, 7:12])
```

SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	A/5 21171	7.2500		S
1	0	PC 17599	71.2833	C85	C
0	0	STON/O2. 3101282	7.9250		S
1	0	113803	53.1000	C123	S
0	0	373450	8.0500		S
0	0	330877	8.4583		Q

El dataset conté les dades de 891 passatgers i 12 variables per cadascún d'ells (el fitxer test només inclou 11 variables):

```
# Número de passatgers
nrow(passatgers)
## [1] 891
# Número de variables
ncol(passatgers)
## [1] 12
```

Aquestes variables són les següents:

```
# Nom de les variables
labels(passatgers)[2]
## [[1]]
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

La primera variable, *PassangerId*, tan sols identifica el passatger amb un identificador numèric seqüencial.
La variable *Survived* indica si el passatger va sobreviure al accident, indicant un 1 si va sobreviure o un 0 en el cas contrari. Aquesta és la variable que no inclou el fitxer test.
La variable *Pclass* indica la classe socio-econòmica del passatger. En aquest cas tenim els valors 1, 2 i 3 corresponents a primera, segona i tercera classe, respectivament.
La variable *Name* conté el nom dels passatger.
La variable *Sex* indica el sexe del passatger. Observant les dades veiem que es descriu amb **male** als homes i **female** a les dones.
La variable *Age* indica l'edat en anys del passatger. En el cas que l'edat sigui menor a 1 any, aquesta serà fraccionada. En el cas que sigui estimada, serà del tipus xx.5.
La variable *SibSp* indica el nombre de germans o cònjugues a bord del Titanic.
La variable *Parch* indica el nombre de pares o fills a bord del Titanic. En el cas que un menor viatgés acompanyat de una cuidadora, el valor d'aquesta variable és 0.
La variable *Ticket* indica el codi del ticket d'embarcament.
La variable *Fare* indica la tarifa que va pagar el passatger.
La variable *Cabin* indica el codi del camarot que ocupava el passatger en l'embarcació.

Finalment, la variable *Embarked* indica el port en el qual pujà a bord del vaixell. Els valors són: **C** = Cherbourg, **Q** = Queenstown i **S** = Southampton.

2. Integració i selecció de les dades d'interès a analitzar

Per tal de realitzar l'anàlisis, ens quedarem amb les variables d'interès per tal de predir la probabilitat de sobreviure. Considerem que aquestes variables seran:

- Survived
- Fare
- Pclass
- Sex
- Age
- SibSp
- Parch

La resta de variables considerem que no son rellevants per tal d'establir una predicció de supervivència. En el cas de les variables *PassangerId*, *Name*, *Ticket*, *Embarked* són identificadors que entenem que no afectaran en la probabilitat de sobreviure.
En el cas de la variable *Cabin*, si realitzem un petit anàlisis observem que conté molts valors perduts.

3. Neteja de les dades

Abans de netejar les dades, procedirem a realitzar una primera revisió d'aquestes.

La variable *Survived* hauria d'estar factoritzada amb els valors 0 i 1. Si realitzem un resum de la variable, obtenim:

```
# Resum de la variable Survived
summary(passatgers$Survived)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3838  1.0000  1.0000
```

Com veiem, no s'està detectant la variable com un factor, per tant, l'haurem de factoritzar.

```
# Factoritzem la variable Survived
passatgers$Survived <- as.factor(passatgers$Survived)
is.factor(passatgers$Survived)
## [1] TRUE
```

Si realitzem un resum de la variable ara, observem com tenim 342 supervivents i 549 no supervivents:

```
# Resum de la variable Survived
summary(passatgers$Survived)
##   0   1
## 549 342
```

Si realitzem un resum de la variable *Fare*, obtenim:

```
# Resum de la variable Fare
summary(passatgers$Fare)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.91   14.45   32.20   31.00  512.33
```

No veiem cap valor NA, però veiem un valor que s'allunya molt de la mitjana i observem zeros. Aquesta casuística la tractarem més endavant.

Si observem la variable *Sex*, observem com tenim 314 dones i 577 homes:

```
# Resum de la variable Sex
summary(passatgers$Sex)
## female   male
##    314    577
```

En aquest cas no hem de realitzar cap tractament d'inici.

Si analitzem la variable *Age*, observem el següent resum:

```
# Resum de la variable Age
summary(passatgers$Age)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

Veiem que apareixen molts valors NA. Aquesta casuística la tractarem més endavant.

Respecte la variable *SibSp*, observem el següent:

```
# Resum de la variable SibSp
summary(passatgers$SibSp)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.523   1.000   8.000
```

En aquest cas, no veiem cap valor NA, però veiem un valor que s'allunya molt de la mitjana i observem zeros. Aquesta casuística la tractarem més endavant.

Per últim, analitzem la variable *Parch*:

```
# Resum de la variable Parch
summary(passatgers$Parch)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3816  0.0000  6.0000
```

Com en el cas anterior, no veiem cap valor NA, però veiem un valor que s'allunya molt de la mitjana i observem zeros.

3.1 Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Hem observat que tenim variables amb elements buits (NA) o que contenen zeros. Procedim a gestionar aquestes casuístiques.

Pel que fa a la variable *Age*, hem vist que conté 177 registres sense l'edat informada. Considerarem que no va nèixer cap nen o nena durant el viatge. Entenem que una edat 0 no pot ser, per tant, per solucionar aquest escenari, imputarem els valors perduts utilitzant el mètode dels k-veïns més propers usant la distància de Gower.

```
# Mètode KNN per imputar valors perduts
passatgers_2 <- kNN(passatgers)[, 1:12]
passatgers_test <- kNN(passatgers_test)[, 1:11]
summary(passatgers_2$Age)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   21.00   28.00   29.43   36.75   80.00
```

Podem observar com ja no tenim cap passatger amb el valor de la variable *Age* buit.

Pel que fa a la variable *Fare*, veiem que conté zeros. Si observem els casos en concret:

```
# Passatges amb tarifa 0
passatgers[which(passatgers$Fare == 0), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
180	180	0	3	Leonard, Mr. Lionel	male	36
264	264	0	1	Harrison, Mr. William	male	40
272	272	1	3	Tornquist, Mr. William Henry	male	25
278	278	0	2	Parkes, Mr. Francis "Frank"	male	NA
303	303	0	3	Johnson, Mr. William Cahoone Jr	male	19
414	414	0	2	Cunningham, Mr. Alfred Fleming	male	NA
467	467	0	2	Campbell, Mr. William	male	NA
482	482	0	2	Frost, Mr. Anthony Wood "Archie"	male	NA
598	598	0	3	Johnson, Mr. Alfred	male	49
634	634	0	1	Parr, Mr. William Henry Marsh	male	NA
675	675	0	2	Watson, Mr. Ennis Hastings	male	NA
733	733	0	2	Knight, Mr. Robert J	male	NA
807	807	0	1	Andrews, Mr. Thomas Jr	male	39
816	816	0	1	Fry, Mr. Richard	male	NA
823	823	0	1	Reuchlin, Jonkheer. John George	male	38

```
passatgers[which(passatgers$Fare == 0), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
180	0	0	LINE	0		S
264	0	0	112059	0	B94	S
272	0	0	LINE	0		S
278	0	0	239853	0		S
303	0	0	LINE	0		S
414	0	0	239853	0		S
467	0	0	239853	0		S
482	0	0	239854	0		S
598	0	0	LINE	0		S
634	0	0	112052	0		S
675	0	0	239856	0		S
733	0	0	239855	0		S
807	0	0	112050	0	A36	S
816	0	0	112058	0	B102	S
823	0	0	19972	0		S

Observem que tots els passatgers tenen un ticket assignat. En el cas dels passatgers amb el ticket 'LINE', després de buscar informació, es tracta de treballadors de l'empresa American Line que van tenir que embarcar al Titanic degut que el vaixell on teniem que embarcar (Philadelphia's) va ser cancel · lat a causa de les vagues del personal del carbó.
Si extreiem, el número de passatgers amb els tickets que tenen els passatgers amb *Fare* igual a 0 o el mateix valor de la variable *Cabin*, veiem que només trobem a ells mateixos:

```
# Passatges amb tarifa 0
nrow(passatgers[which(passatgers_2$Ticket == "LINE"), ])
## [1] 4
nrow(passatgers[which(passatgers_2$Ticket == 112059), ])
## [1] 1
nrow(passatgers[which(passatgers_2$Cabin == "B94"), ])
## [1] 1
nrow(passatgers[which(passatgers_2$Ticket == 239853), ])
```

```
## [1] 3
nrow(passatgers_2[which(passatgers_2$Ticket == 239854), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Ticket == 112052), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Ticket == 239856), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Ticket == 239855), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Ticket == 112050), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Cabin == "A36"), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Ticket == 112058), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Cabin == "B102"), ])
## [1] 1
nrow(passatgers_2[which(passatgers_2$Ticket == 19972), ])
## [1] 1
```

Per tant, per tal d'imputar els valors de la variable *Fare* d'aquests passatgers, hem decidit imputar la mitjana de la variable segons la variable *Pclass*:

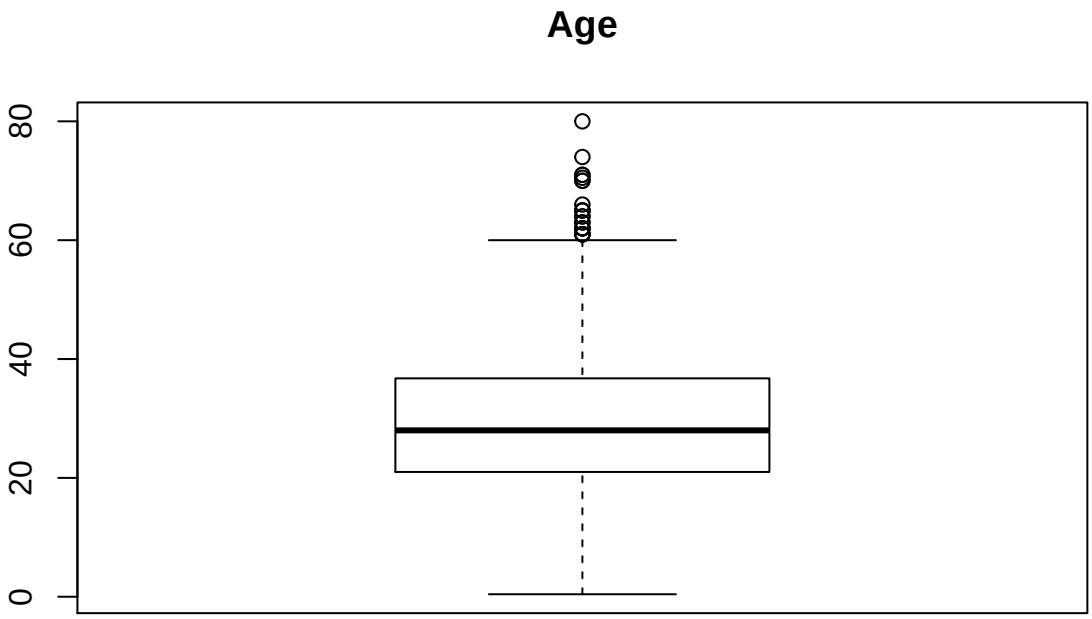
```
passatgers_2[which(passatgers_2$Fare == 0 & passatgers_2$Pclass == 3), "Fare"] <- mean(passatgers_2[whi
    3), "Fare"])
passatgers_2[which(passatgers_2$Fare == 0 & passatgers_2$Pclass == 2), "Fare"] <- mean(passatgers_2[whi
    2), "Fare"])
passatgers_2[which(passatgers_2$Fare == 0 & passatgers_2$Pclass == 1), "Fare"] <- mean(passatgers_2[whi
    1), "Fare"])
```

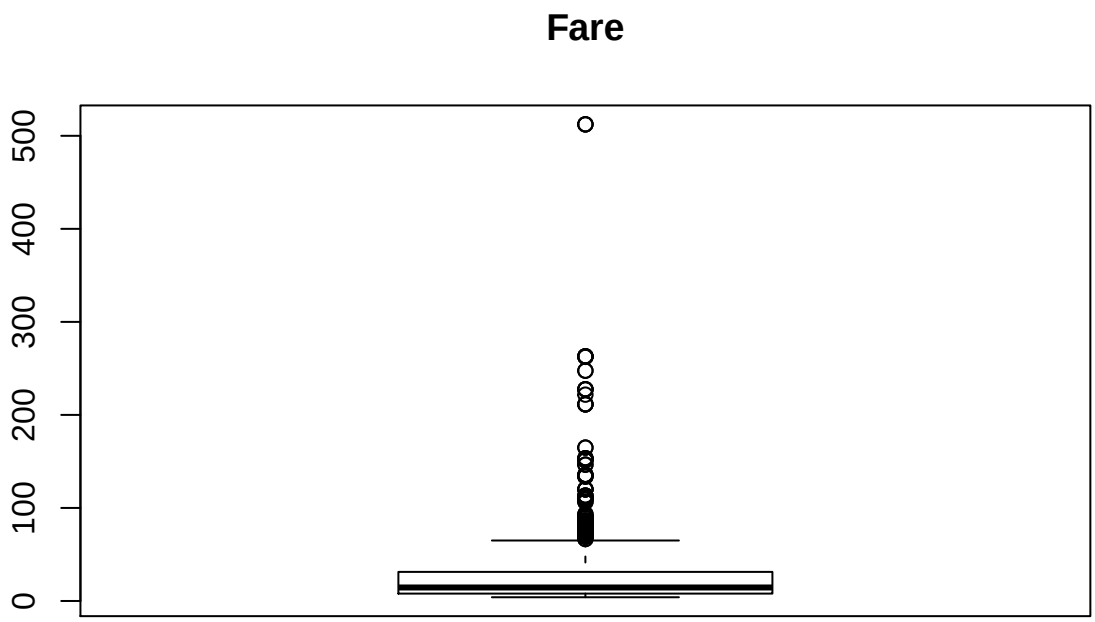
Pel que fa a les variables *SibSp* i *Parch*, contenen valors zeros. En el cas de la variable *SibSp* és una casuística plausible degut que un passatger podria viatgar sense germans ni cònjugues (fill únic i solter, o simplement viatja sol). En el cas de la variable *Parch* abans hem comentat que si un menor viatjava acompanyat de una cuidadora, el valor d'aquesta variable és 0.

3.2 Identificació i tractament de valors extrems

Per tal d'identificar possibles valors extrems, mostrarem gràfics de caixa per les variables *Age* i *Fare*:

```
# Boxplots Age, SibSp i Parch
variables <- c("Age", "Fare")
for (name in variables) {
    boxplot(passatgers_2[, name], main = name)
}
```


Age
0
20
40
60
80



Els extrems que es mostren en els gràfics són el següents:

```
# Valors extrems
boxplot.stats(passatgers_2$Age)$out
##  [1] 66.0 65.0 71.0 70.5 61.0 61.0 62.0 63.0 65.0 61.0 64.0 65.0 63.0 71.0
## [15] 64.0 62.0 62.0 61.0 61.0 80.0 70.0 70.0 62.0 74.0
boxplot.stats(passatgers_2$Fare)$out
##   [1]  71.28330 263.00000 146.52080  82.17080  76.72920  80.00000  83.47500
##   [8]  73.50000 263.00000  77.28750 247.52080  73.50000  77.28750  79.20000
##  [15]  66.60000  69.55000  69.55000 146.52080  69.55000 113.27500  76.29170
##  [22]  90.00000  83.47500  90.00000  79.20000  86.50000 512.32920  79.65000
##  [29]  84.15469 153.46250 135.63330  77.95830  78.85000  91.07920 151.55000
##  [36] 247.52080 151.55000 110.88330 108.90000  83.15830 262.37500 164.86670
##  [43] 134.50000  69.55000 135.63330 153.46250 133.65000  66.60000 134.50000
##  [50] 263.00000  75.25000  69.30000 135.63330  82.17080 211.50000 227.52500
##  [57]  73.50000 120.00000 113.27500  90.00000 120.00000 263.00000  81.85830
##  [64]  89.10420  91.07920  90.00000  78.26670 151.55000  86.50000 108.90000
##  [71]  93.50000 221.77920 106.42500  71.00000 106.42500 110.88330 227.52500
##  [78]  79.65000 110.88330  79.65000  79.20000  78.26670 153.46250  77.95830
##  [85]  84.15469  69.30000  76.72920  73.50000 113.27500 133.65000  73.50000
##  [92] 512.32920  76.72920 211.33750 110.88330 227.52500 151.55000 227.52500
##  [99] 211.33750 512.32920  78.85000 262.37500  71.00000  86.50000 120.00000
## [106]  77.95830 211.33750  79.20000  69.55000 120.00000  84.15469  84.15469
## [113]  93.50000  84.15469  80.00000  83.15830  69.55000  89.10420 164.86670
## [120]  69.55000  83.15830
```

Podem observar que el valors extrems de la variable *Age* entren dins de la realitat.

En el cas de la variable *Fare*, veiem molts valors extrems. Si realitzem un revisió per damunt d'aquest casos tenim que la majoria de passatgers són de primera classe:

```
head(passatgers[which(passatgers$Fare >= min(boxplot.stats(passatgers_2$Fare)$out)),
    1:6])
```

	PassengerId	Survived	Pclass	Name	Sex	Age
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
28	28	0	1	Fortune, Mr. Charles Alexander	male	19
32	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NA
35	35	0	1	Meyer, Mr. Edgar Joseph	male	28
53	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49
62	62	1	1	Icard, Miss. Amelie	female	38

```
head(passatgers[which(passatgers$Fare >= min(boxplot.stats(passatgers_2$Fare)$out)),
    7:12])
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	PC 17599	71.2833	C85	C
28	3	2	19950	263.0000	C23 C25 C27	S
32	1	0	PC 17569	146.5208	B78	C
35	1	0	PC 17604	82.1708		C
53	1	0	PC 17572	76.7292	D33	C
62	0	0	113572	80.0000	B28	

Però si observem més detalladament trobaríem casos de tercera i segona classe:

```
passatgers[which(passatgers$Fare >= min(boxplot.stats(passatgers_2$Fare)$out) & passatgers$Pclass ==
    3), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
160	160	0	3	Sage, Master. Thomas Henry	male	NA
181	181	0	3	Sage, Miss. Constance Gladys	female	NA
202	202	0	3	Sage, Mr. Frederick	male	NA
325	325	0	3	Sage, Mr. George John Jr	male	NA
793	793	0	3	Sage, Miss. Stella Anna	female	NA
847	847	0	3	Sage, Mr. Douglas Bullen	male	NA
864	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NA

```
passatgers[which(passatgers$Fare >= min(boxplot.stats(passatgers_2$Fare)$out) & passatgers$Pclass ==
    3), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
160	8	2	CA. 2343	69.55		S
181	8	2	CA. 2343	69.55		S
202	8	2	CA. 2343	69.55		S
325	8	2	CA. 2343	69.55		S
793	8	2	CA. 2343	69.55		S
847	8	2	CA. 2343	69.55		S
864	8	2	CA. 2343	69.55		S

```
passatgers[which(passatgers$Fare >= min(boxplot.stats(passatgers_2$Fare)$out) & passatgers$Pclass ==
    2), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
73	73	0	2	Hood, Mr. Ambrose Jr	male	21
121	121	0	2	Hickman, Mr. Stanley George	male	21
386	386	0	2	Davies, Mr. Charles Henry	male	18
656	656	0	2	Hickman, Mr. Leonard Mark	male	24
666	666	0	2	Hickman, Mr. Lewis	male	32

```
passatgers[which(passatgers$Fare >= min(boxplot.stats(passatgers_2$Fare)$out) & passatgers$Pclass ==
    2), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
73	0	0	S.O.C. 14879	73.5		S
121	2	0	S.O.C. 14879	73.5		S
386	0	0	S.O.C. 14879	73.5		S
656	2	0	S.O.C. 14879	73.5		S
666	2	0	S.O.C. 14879	73.5		S

Veiem com es tracta de casos en que el ticket d'embarcament es compartit entre varies persones, la qual cosa fa pujar la seva tarifa.

Si realitzem una agrupació per factors de la variable *SibSp*, observem els següents factors (en aquest cas unirem els fitxer train i test per tal de tenir tots els passtagers):

```
passatgers_test$Survived <- NaN
passatgers_test <- subset(passatgers_test, select = c(1, 12, 2, 3, 4, 5, 6, 7, 8,
    9, 10, 11))
passatgers_all <- rbind(passatgers, passatgers_test)
nrow(passatgers_all)
## [1] 1309
grupos_sibsp <- group_by(passatgers_all, passatgers_all$SibSp)
## Warning: package 'bindrcpp' was built under R version 3.4.4
resum_grupos_sibsp <- summarise(grupos_sibsp, num = n())
resum_grupos_sibsp
```

passatgers_all$SibSp	num
0	891
1	319
2	42
3	20
4	22
5	6
8	9

Podem observar uns valors candidats a ser erronis. Si analitzem els passatgers amb valor 8:

```
# Passatgers amb 8 germans i cònjugues
passatgers_all[which(passatgers_all$SibSp == 8), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
160	160	0	3	Sage, Master. Thomas Henry	male	NA
181	181	0	3	Sage, Miss. Constance Gladys	female	NA
202	202	0	3	Sage, Mr. Frederick	male	NA
325	325	0	3	Sage, Mr. George John Jr	male	NA
793	793	0	3	Sage, Miss. Stella Anna	female	NA
847	847	0	3	Sage, Mr. Douglas Bullen	male	NA
864	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NA
1080	1080	NA	3	Sage, Miss. Ada	female	14.5
1252	1252	NA	3	Sage, Master. William Henry	male	14.5

```
passatgers_all[which(passatgers_all$SibSp == 8), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
160	8	2	CA. 2343	69.55		S
181	8	2	CA. 2343	69.55		S
202	8	2	CA. 2343	69.55		S
325	8	2	CA. 2343	69.55		S
793	8	2	CA. 2343	69.55		S
847	8	2	CA. 2343	69.55		S
864	8	2	CA. 2343	69.55		S
1080	8	2	CA. 2343	69.55		S
1252	8	2	CA. 2343	69.55		S

Podem observar que es tracta de 9 germans, és a dir, en principi la dada *SibSp* és correcta.

Si observem el valor extrem 5 per la variable *SibSp* i fem les mateixes comprovacions, veurem que el valor també és correcte:

```
# Passatgers amb 5 germans i cònjugues
passatgers_all[which(passatgers_all$SibSp == 5), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
60	60	0	3	Goodwin, Master. William Frederick	male	11
72	72	0	3	Goodwin, Miss. Lillian Amy	female	16
387	387	0	3	Goodwin, Master. Sidney Leonard	male	1
481	481	0	3	Goodwin, Master. Harold Victor	male	9
684	684	0	3	Goodwin, Mr. Charles Edward	male	14
1032	1032	NA	3	Goodwin, Miss. Jessie Allis	female	10

```
passatgers_all[which(passatgers_all$SibSp == 5), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
60	5	2	CA 2144	46.9		S
72	5	2	CA 2144	46.9		S
387	5	2	CA 2144	46.9		S
481	5	2	CA 2144	46.9		S
684	5	2	CA 2144	46.9		S
1032	5	2	CA 2144	46.9		S

Si observesim els altres valor de *SibSp*, ens trobaríem que els valors també són correctes.

Pel que fa a la variable *Parch*, realitzem el mateix anàlisis:

```
# Valors extrems de Parch
grupos_parch <- group_by(passatgers_all, passatgers_all$Parch)
resum_grupos_parch <- summarise(grupos_parch, num = n())
resum_grupos_parch
```

passatgers_all$Parch	num
0	1002
1	170
2	113
3	8
4	6
5	6
6	2
9	2

Ara veiem que els valors extrems són 4, 5, 6 i 9. Fem un anàlisis del valor 9 i 6:

```
# Passatgers amb 9 pares o fills
passatgers_all[which(passatgers_all$Parch == 9), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
1234	1234	NA	3	Sage, Mr. John George	male	19
1257	1257	NA	3	Sage, Mrs. John (Annie Bullen)	female	19

```
passatgers_all[which(passatgers_all$Parch == 9), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1234	1	9	CA. 2343	69.55		S
1257	1	9	CA. 2343	69.55		S

```
# Passatgers amb 6 pares o fills
passatgers_all[which(passatgers_all$Parch == 6), 1:6]
```

	PassengerId	Survived	Pclass	Name	Sex	Age
679	679	0	3	Goodwin, Mrs. Frederick (Augusta Tyler)	female	43
1031	1031	NA	3	Goodwin, Mr. Charles Frederick	male	40

```
passatgers_all[which(passatgers_all$Parch == 6), 7:12]
```

	SibSp	Parch	Ticket	Fare	Cabin	Embarked
679	1	6	CA 2144	46.9		S
1031	1	6	CA 2144	46.9		S

Observem que corresponen als pares dels germans 'Sage' i 'Goodwin' respectivament.

Si observesim els altres valors de *Parch*, ens trobaríem que els valors també són correctes.

4. Anàlisi de les dades

L'objectiu que ens hem plantejat és estimar un model de regressió logística amb variable dependent *Survived*.

4.1 Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

En primer lloc, recodifiquem la variable *Sex* de la següent forma:

```
passatgers_2$Sex <- recode(passatgers_2$Sex, "'male'='M';'female'='F'")
summary(passatgers_2$Sex)
##   F   M
## 314 577
```

A continuació creem la variable nova *SexR* reordenada amb el valor de referència **F**:

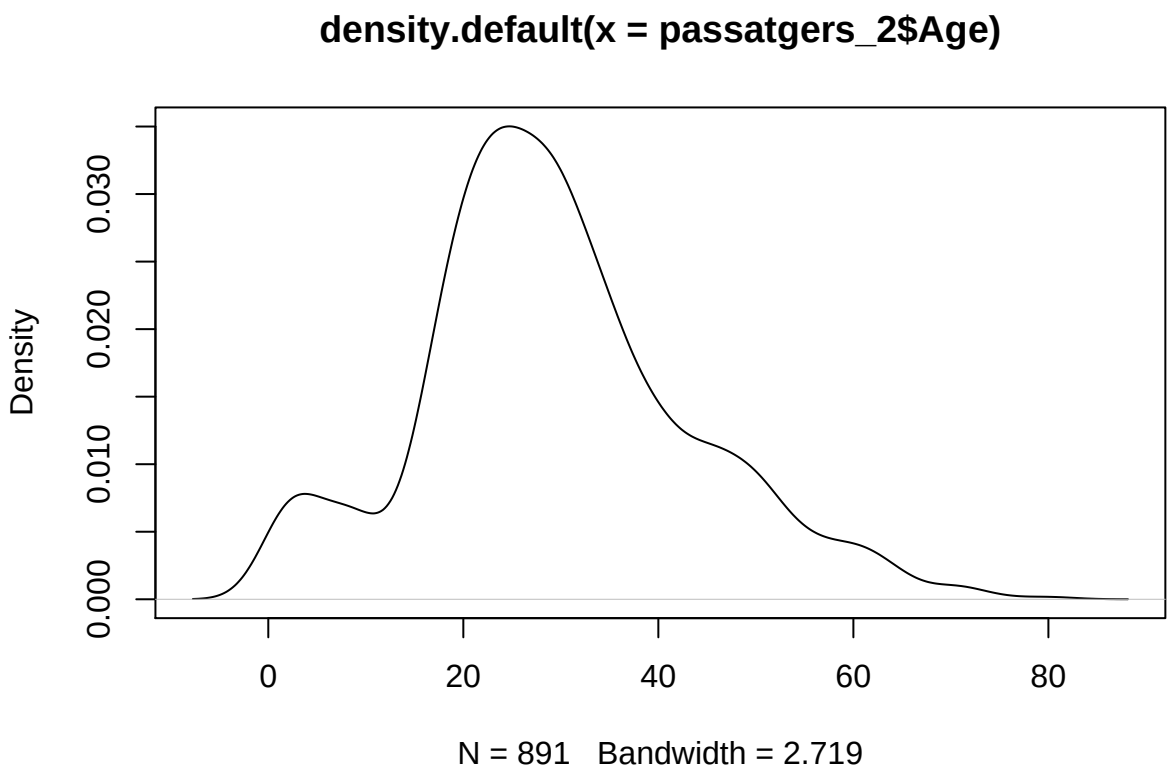
```
# Reordenem la variable Sex en la nova variable SexR
passatgers_2$SexR <- relevel(passatgers_2$Sex, ref = "F")
head(passatgers_2$SexR)
## [1] M F F F M M
## Levels: F M
```

4.2 Comprovació de la normalitat i homogeneïtat de la variància.

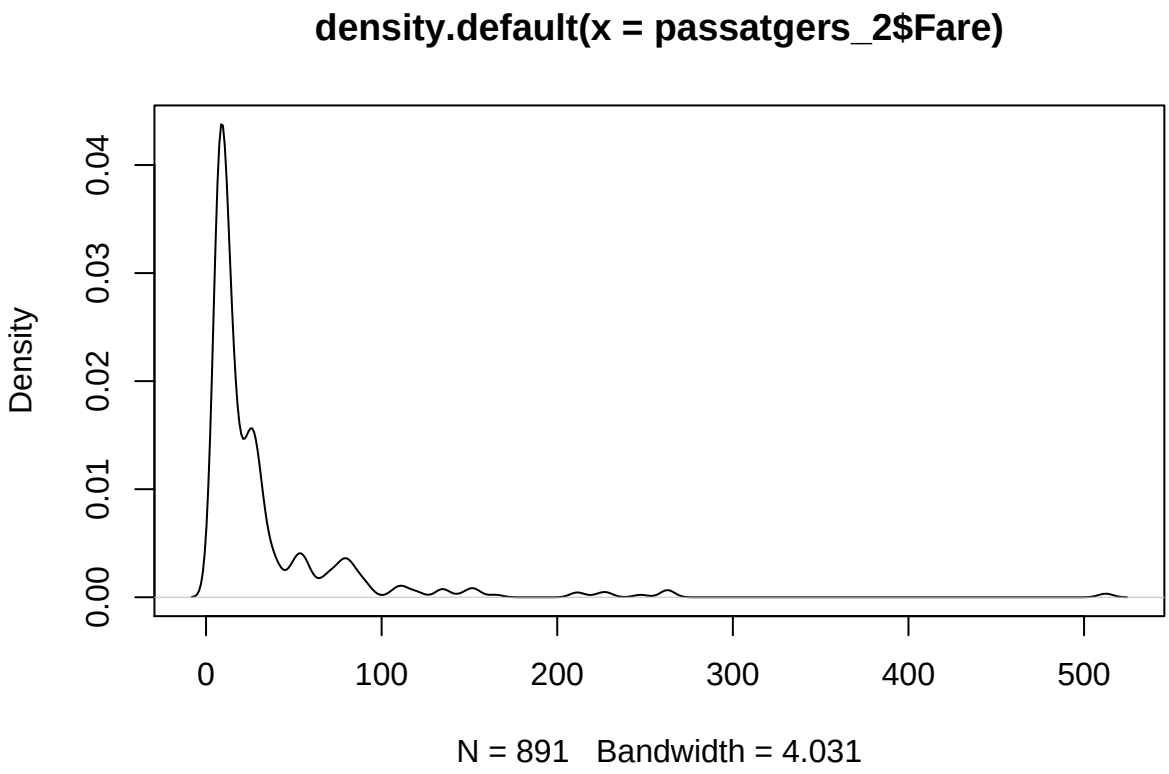
Abans de generar el model de regressió logística, realitzem comprobacions de normalitat i homogeneïtat de la variància.

Si mostrem les gràfiques de les variables *Age* i *Fare*, observem que no tenen una densidad poblacional gaire normal. Si apliquem el test de Pearson, veurem que el valor p de la mostra és molt petit, per tant hauríem de rebutjar la hipòtesis que la mostra te una distribció normal.

```
# Normalitat de Age i Fare
library(nortest)
plot(density(passatgers_2$Age))
```




```
pearson.test(passatgers_2$Age)
##
##  Pearson chi-square normality test
##
## data:  passatgers_2$Age
## P = 181.34, p-value < 2.2e-16
plot(density(passatgers_2$Fare))
```




```
pearson.test(passatgers_2$Fare)
##
##  Pearson chi-square normality test
##
## data:  passatgers_2$Fare
## P = 3712.2, p-value < 2.2e-16
```

Respecte la homogeneïtat, podem aplicar el test de Bartlett i obtenim els resultats següents:

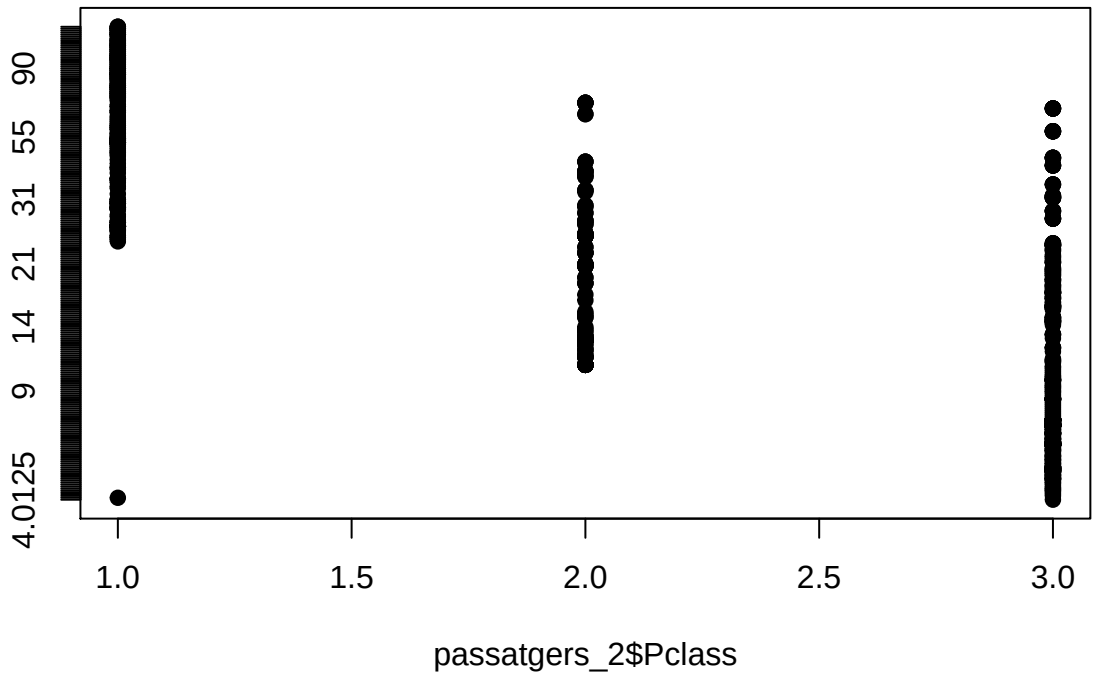
```
# Homogeneïtat de Age i Fare
bartlett.test(passatgers_2$Age ~ passatgers_2$Survived, data = passatgers_2)
##
##  Bartlett test of homogeneity of variances
##
## data:  passatgers_2$Age by passatgers_2$Survived
## Bartlett's K-squared = 0.96519, df = 1, p-value = 0.3259
bartlett.test(passatgers_2$Fare ~ passatgers_2$Survived, data = passatgers_2)
##
##  Bartlett test of homogeneity of variances
##
## data:  passatgers_2$Fare by passatgers_2$Survived
## Bartlett's K-squared = 236.59, df = 1, p-value < 2.2e-16
```

Podem observar com la variable *Age* ens dona una bona homogeneïtat en les dues classes de la variable *Survived*, però veiem que la variable *Fare* no.

4.3 Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

Per poder escollir els regressors del model que volem construir, abans mirarem la correlació entre les variables *Fare* i *Pclass*:

```
# Correlació entre Pclass i Fare
stripchart(passatgers_2$Pclass ~ passatgers_2$Fare, pch = 19)
```




```
cor(passatgers_2$Fare, passatgers_2$Pclass)
## [1] -0.5644407
```

Veiem que el resultat té un valor absolut alt, és a dir, que existeix una correlació entre les variables. El signe negatiu indica que quan més gran és el valor de *Fare*, més baix és el "valor" de *Pclass*. Com que els "valors" de *Pclass* són 1 = primera classe, 2 = classe mitjana i 3 = classe baixa, podem dir que els passatgers de primera classe han pagat tarifes superiors a les altres classes.

Com s'observa en el gràfic, no inclourem la variable *Pclass*, degut que existeix una dependència entre les variables *Pclass* i *Fare*. Això implica un problema de multicol · linealitat, la qual cosa ocasiona efectes molt importants en les estimacions i els resultats poden ser confusos.

Per tant, el fitxer resultat amb el qual realitzarem l'analisi és Titanic_clean.csv:

```
# Escriptura del fitxer netejat
passatgers_clean <- cbind(passatgers_2[, 1:4], passatgers_2$SexR, passatgers_2[,
    6:12])
write.csv(passatgers_clean, file = paste(ruta, "Titanic_clean.csv", sep = ""), sep = ",")
```

Si apliquem el model amb un nivell de significància del 0,05, obtenim el següent:

```
# Generem el model
model_logit <- glm(passatgers_2$Survived ~ passatgers_2$Fare + passatgers_2$SexR +
    passatgers_2$Age + passatgers_2$SibSp + passatgers_2$Parch, family = "binomial")
summary(model_logit)
##
## Call:
## glm(formula = passatgers_2$Survived ~ passatgers_2$Fare + passatgers_2$SexR +
##     passatgers_2$Age + passatgers_2$SibSp + passatgers_2$Parch,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5115  -0.6582  -0.5285   0.7165   2.3831
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.623246   0.262835   6.176 6.58e-10 ***
## passatgers_2$Fare    0.016044   0.002761   5.810 6.25e-09 ***
## passatgers_2$SexRM -2.615310   0.186277 -14.040  < 2e-16 ***
## passatgers_2$Age   -0.023582   0.006924  -3.406  0.00066 ***
## passatgers_2$SibSp -0.440882   0.103239  -4.271 1.95e-05 ***
## passatgers_2$Parch -0.228285   0.112862  -2.023  0.04311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  850.51  on 885  degrees of freedom
## AIC: 862.51
##
## Number of Fisher Scoring iterations: 5
```

Veiem que tots els regressors són significants.

Amb aquest model, els coeficients de regressió són els següents:

```
# Coeficients de regressió
coeff_beta <- coefficients(model_logit)
coeff_beta
##        (Intercept)  passatgers_2$Fare passatgers_2$SexRM
##         1.62324609         0.01604428        -2.61531018
##   passatgers_2$Age passatgers_2$SibSp passatgers_2$Parch
##        -0.02358171        -0.44088153        -0.22828498
```

Per tant el model de regressió logística és:

$$Prob(Y_i = 1) = \frac{exp(\beta_0 + \beta_1 Fare_i + \beta_2 SexR_i + \beta_3 Age_i + \beta_4 SibSp_i + \beta_5 Parch_i)}{1 + exp(\beta_0 + \beta_1 Fare_i + \beta_2 SexR_i + \beta_3 Age_i + \beta_4 SibSp_i + \beta_5 Parch_i)}$$

Segons els valors obtinguts podem realitzar els anàlisis següents:

- El coeficient d'intersecció no té sentit analitzar-lo degut que les variables *Age* i *Fare* no poden ser 0.
- Si l'individu va pagar una tarifa (*Fare*) alta, és a dir, és de primera classe, té més probabilitats de sobreviure.

- Si l'individu és dona, la probabilitat de sobreviure depèn de la resta de variables *Fare*, *Age*, *SibSp* i *Parch*.
- Si l'individu és home, la probabilitat de sobreviure disminueix.
- Com més edat tingui l'individu, menys probabilitat de sobreviure té.
- Com més germans i/o cónjugues tingués a bord, menys probabilitat de sobreviure té.
- Com més pares i/o fills tingués a bord, menys probabilitat de sobreviure té.

Passem a anlitzar la qualitat d'ajust del model creat. En primer lloc, creem un dataframe on la primera columna sigui les observacións dels nostre conjunt de dades si un individu sobreviu o no, és a dir, la variable *Survived*, i la segona columna sigui els valors predits pel model anterior amb un llindar de discriminació del 70%:

```
# Calculem els valors predits
valors_predits <- predict(model_logit, passatgers_2, type = "response")
head(valors_predits)
##         1         2         3         4         5         6
## 0.1375936 0.8068934 0.7571785 0.7701142 0.1560038 0.2056134
# Interpretem els resultats amb el llindar indicat
clase_predita <- ifelse(valors_predits > 0.7, 1, 0)
head(clase_predita)
## 1 2 3 4 5 6
## 0 1 1 1 0 0
# Montem el data set a analitzar
data <- data.frame(obs = passatgers_2$Survived, pre = clase_predita)
kable(data.frame(Observació = head(data$obs), Predicció = head(data$pre)), align = c("l",
    "l"))
```

Observació	Predicció
0	0
1	1
1	1
1	1
0	0
0	0

5. Representació dels resultats a partir de taules i gràfiques.

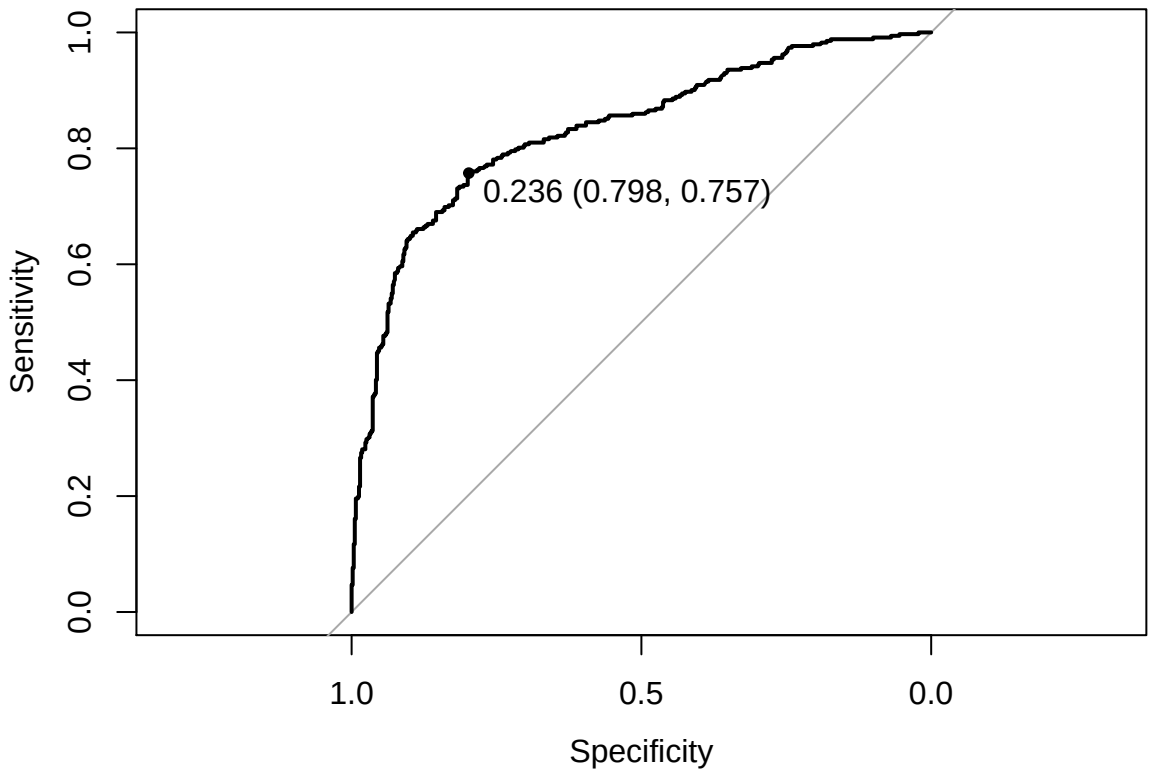
Amb aquests resultats obtinguts, podem montar una taula de doble entrada amb cadascuna de les variables anterior i obtenim la matriu de confusió.

```
# Montem la matriu de confusió
matriu_confusio <- table(data$obs, data$pre, dnn = c("Observació", "Predicció"))
matriu_confusio
##           Predicció
## Observació   0   1
##          0 510  39
##          1 150 192
```

Podem veure com tenim 39 falsos positius i 150 falsos negatius. Els falsos positius són aquells, en el nostre cas, que no sobreviuen i el model ha predit que si. En contra, els falsos negatius són aquells que sobreviuen i el model a predit que no.

Per tal de mostrar la qualitat del model, podem mostrar la corba ROC associada:

```r
# Calculem la corba ROC
roc <- roc(passatgers_2$Survived, valors_predits)
# Mostrem la corba calculada en un gràfic
plot(roc, print.thres = "best", print.thres.best.method = "closest.topleft")
```



Com podem veure, la forma de la corba s'aproxima molt a la cantonada superiror-esquerra del gràfic, la qual cosa indica que la qualitat del model és alta. Una altra dada que és pot extreure del gràfic anterior és el llindar òptim, és a dir, el llindar de discriminació que maximitza la sensibilitat i la especificitat del model, o amb altres paraules, el que minimitza els falsos positius i falsos negatius.

Si apliquem aquest llindar al model, obtenim la matriu de confusió següent:

```r
clase_predita_2 <- ifelse(valors_predits > 0.236, 1, 0)
# Montem el data set a analitzar
data_2 <- data.frame(obs_2 = passatgers_2$Survived, pre_2 = clase_predita_2)
# Montem la matriu de confusió
matriu_confusio_2 <- table(data_2$obs_2, data_2$pre_2, dnn = c("Observació", "Predicció"))
matriu_confusio_2
##           Predicció
## Observació   0   1
##          0 438 111
##          1  83 259
```

En les següents taules mostrem els valors interessants que es poden extreure de la matriu de confusió:

```r
# Valors descriptius de la predicció
positius <- sum(data_2$obs_2 == 1)
negatius <- sum(data_2$obs_2 == 0)
```

```r
positius_predit <- sum(data_2$pre_2 == 1)
negatius_predit <- sum(data_2$pre_2 == 0)
total <- nrow(data_2)
kable(data.frame(Mesura = c("Positius", "Negatius", "Positius predits", "Negatius Predits"),
    Valor = c(positius, negatius, positius_predit, negatius_predit)), align = c("l",
    "l", "l", "l"))
```

Mesura	Valor
Positius	342
Negatius	549
Positius predits	370
Negatius Predits	521

```r
tp <- sum(data_2$obs_2 == 1 & data_2$pre_2 == 1)
tn <- sum(data_2$obs_2 == 0 & data_2$pre_2 == 0)
fp <- sum(data_2$obs_2 == 0 & data_2$pre_2 == 1)
fn <- sum(data_2$obs_2 == 1 & data_2$pre_2 == 0)
kable(data.frame(Mesura = c("Certs positius", "Certs negatius", "Falsos positius",
    "Falsos negatius"), Valor = c(tp, tn, fp, fn)), align = c("l", "l", "l", "l"))
```

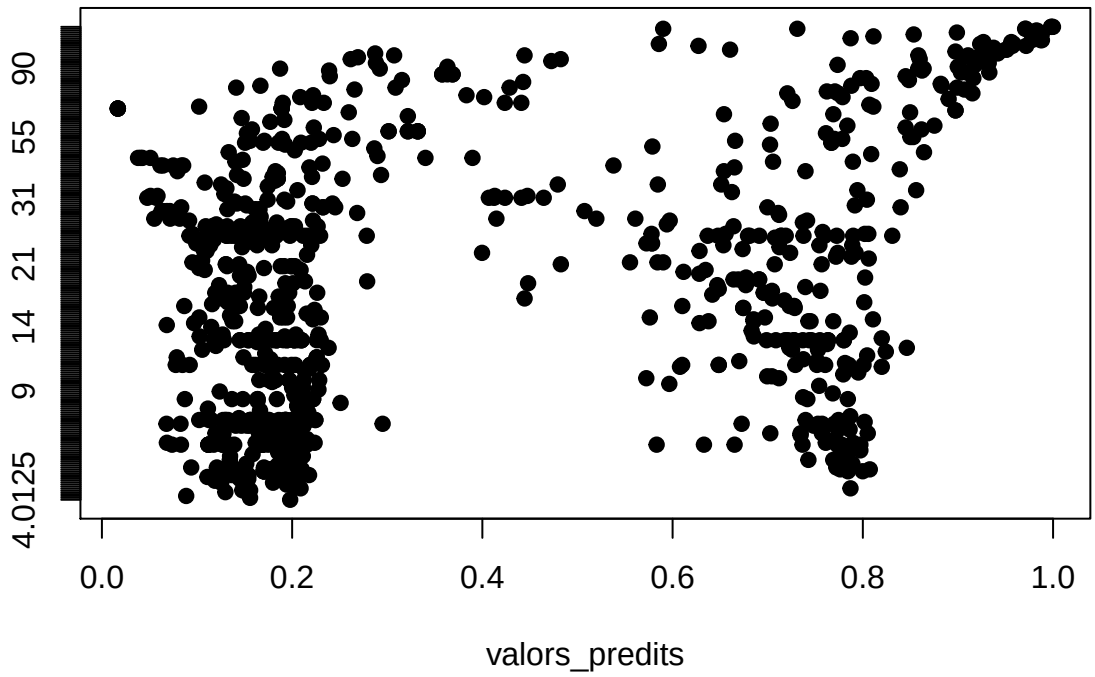
Mesura	Valor
Certs positius	259
Certs negatius	438
Falsos positius	111
Falsos negatius	83

```r
exactitut <- (tp + tn)/total
ratio_error <- (fp + fn)/total
sensibilitat <- tp/positius
especificitat <- tn/negatius
precisio <- tp/positius_predit
valor_pre_neg <- tn/negatius_predit
kable(data.frame(Mesura = c("Exactitud", "Ratio d'error", "Sensibilitat", "Especificitat",
    "Precisió", "Valor de predicció de negatius"), Valor = c(exactitut, ratio_error,
    sensibilitat, especificitat, precisio, valor_pre_neg)), align = c("l", "l", "l",
    "l", "l", "l"))
```

Mesura	Valor
Exactitud	0.7822671
Ratio d'error	0.2177329
Sensibilitat	0.7573099
Especificitat	0.7978142
Precisió	0.7000000
Valor de predicció de negatius	0.8406910

Ara mostrem un gràfic amb la comparació de la probabilitat de sobreviure amb la tarifa que varen pagar dels individus:

```
stripchart(valors_predits ~ passatgers_2$Fare, pch = 19)
```



Com podem veure en aquest gràfic, hi ha una lleugera tendència a augmentar la probabilitat de sobreviure quan la tarifa pagada és superior. Però observem casos de tarifa baixa amb una probabilitat alta de sobreviure. Si extreiem els individus amb una tarifa baixa amb una probabilitat superior al 70% de sobreviure, trobem el següent:

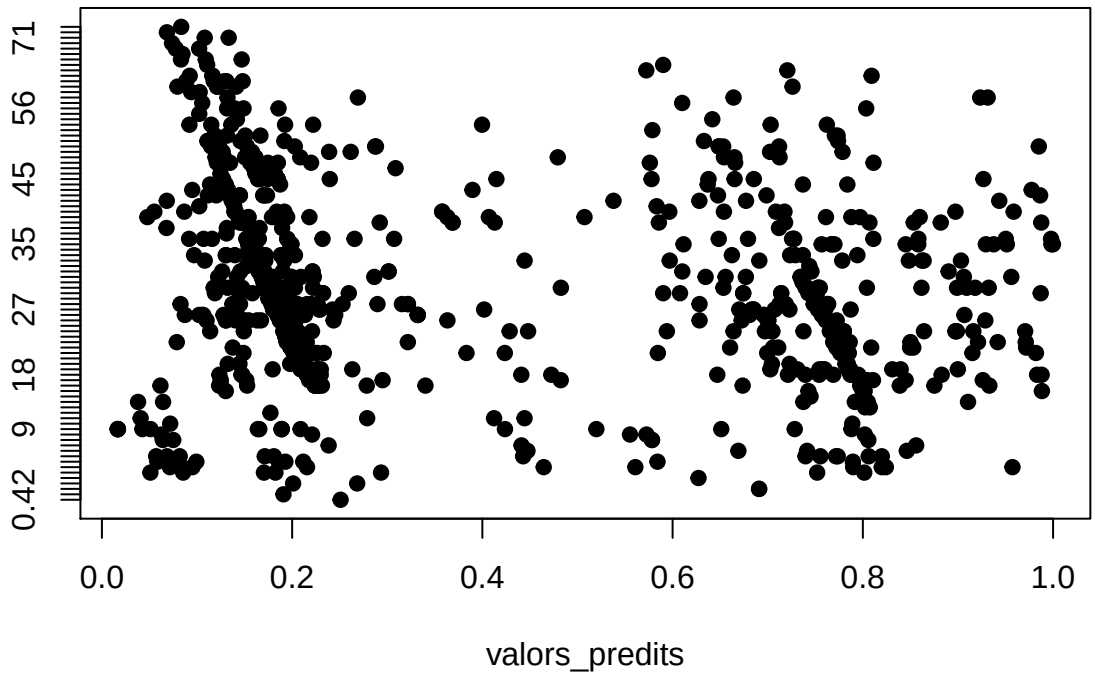
```
passatgers_2$Ppre <- valors_predits
passatgers_2$Pre <- data_2$pre_2
p_low_sur_pre <- passatgers_2[which(passatgers_2$Ppre >= 0.7 & passatgers_2$Fare <=
    quantile(passatgers_2$Fare)[2]), ]
kable(head(p_low_sur_pre[4:7]))
```

	Name	Sex	Age	SibSp
3	Heikkinen, Miss. Laina	F	26	0
15	Vestrom, Miss. Hulda Amanda Adolfina	F	14	0
20	Masselmani, Mrs. Fatima	F	19	0
29	O'Dwyer, Miss. Ellen "Nellie"	F	26	0
33	Glynn, Miss. Mary Agatha	F	22	0
45	Devaney, Miss. Margaret Delia	F	19	0

Podem observar que tots els individus són dones amb una edat baixa. Com hem dit, el fet de ser dona fa pujar la probabilitat de sobreviure, com també l'augmenta tenir una edat baixa.

Si mostrem un gràfic amb la comparació de la probabilitat de sobreviure amb la edat dels individus:

```
stripchart(valors_predits ~ passatgers_2$Age, pch = 19)
```



Com podem veure en aquest gràfic, hi ha una lleugera tendència a augmentar la probabilitat de sobreviure quan la edat és menor. Però observem casos de baixa edat amb poca probabilitat de sobreviure i casos de edad avançada amb alta probabilitat.
Si extreiem els individus de baixa edat (primer quartil) amb una probabilitat inferior al 20% de sobreviure, trobem el següent:

```
p_lage_sur_pre <- passatgers_2[which(passatgers_2$Ppre <= 0.2 & passatgers_2$Age <=
    quantile(passatgers_2$Age)[2]), ]
kable(head(p_lage_sur_pre[3:7]))
```

	Pclass	Name	Sex	Age	SibSp
8	3	Palsson, Master. Gosta Leonard	M	2	3
17	3	Rice, Master. Eugene	M	2	4
51	3	Panula, Master. Juha Niilo	M	7	4
60	3	Goodwin, Master. William Frederick	M	11	5
64	3	Skoog, Master. Harald	M	4	3
66	3	Moubarek, Master. Gerios	M	4	1

Podem observar que la majoria dels individus són homes i tots de classe baixa. Com hem dit, el fet de ser home disminueix la probabilitat de sobreviure, com també pertànyer a la tercera classe (*Fare* baixa).

Si extreiem els individus de edat avançada (tercer quartil) amb una probabilitat superior al 70% de sobreviure, trobem el següent:

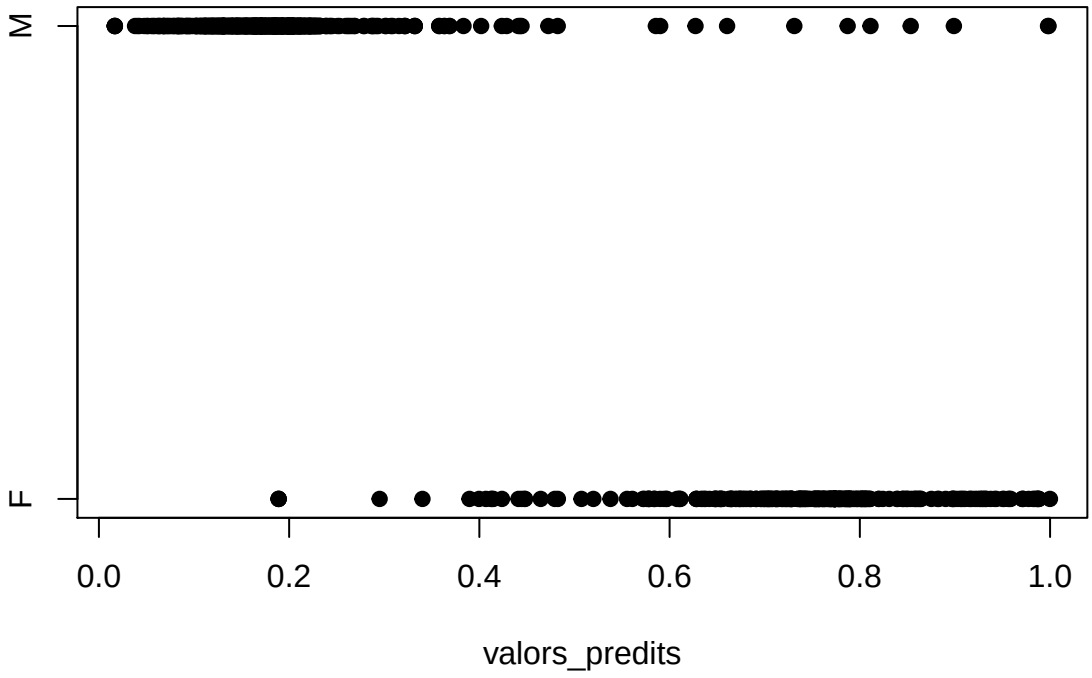
```
p_hage_sur_pre <- passatgers_2[which(passatgers_2$Ppre >= 0.7 & passatgers_2$Age >=
    quantile(passatgers_2$Age)[4]), ]
kable(head(p_hage_sur_pre[3:7]))
```

	Pclass	Name	Sex	Age	SibSp
2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	F	38	1
53	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	F	49	1
62	1	Icard, Miss. Amelie	F	38	0
162	2	Watt, Mrs. James (Elizabeth "Bessie" Inglis Milne)	F	40	0
178	1	Isham, Miss. Ann Elizabeth	F	50	0
195	1	Brown, Mrs. James Joseph (Margaret Tobin)	F	44	0

Podem observar que la majoria dels individus són dones de classe alta (*Fare* alta).

Per últim, si mostrem un gràfic amb la comparació de la probabilitat de sobreviure amb el sexe dels individus:

```
stripchart(valors_predits ~ passatgers_2$SexR, pch = 19)
```



Veiem clarament, que la probabilitat de sobreviure és major en els casos de les dones que en els casos dels homes.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Després de dur a terme l'anàlisi de les dades del fitxer Titanic.csv, i partint de la pregunta a resoldre de si existeix algun grup de passatgers amb més probabilitats de sobreviure a l'accident, podem concloure que les possiblitats de sobreviure augmenten segons l'edat, el sexe i la classe del passatger. En altres pasarules, les dones i els nens de primera classe seria un dels grups amb més possibilitat de sobreviure. En menys mesura, les unitats familiars més petites també seria un grup que podria augmentar aquesta possibilitat.

Per tant, podríem aventurar que els responsables de la tripulació destinada a pujar a bord dels bots salvavides, el dia del naufragi, varen decidir prioritzar salvar la vida als nens i a les dones que els acompanyaven. El fet que els passatgers de primera classe tinguessin més possibilitats de sobreviure, pot ser degut que els camarots destinats a la tercera classe es trobaven en la part inferior del vaixell, i un cop arribaren a coberta, en el cas d'aconsegui-ho, ja no quedaven suficients bots per la gent que esperava per pujar en algun.